AD 290 313

ESD-TR-62-294

STUDIES FOR THE DESIGN OF AN ENGLISH COMMAND AND CONTROL LANGUAGE SYSTEM

Arthur D. Little, Inc.
35 Acorn Park, Cambridge, Massachusetts

Report Number CACL-2

Contract No. AF 19(628)-256

November 1962

Prepared
for

OPERATIONAL APPLICATIONS LABORATORY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS

Requests for additional copies by Agencies of the Department of Defense, their contractors, and other Government agencies should be directed to the:

ARMED SERVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA

Department of Defense contractors must be established for ASTIA services or have their 'need-to-know' certified by the cognizant military agency of their project or contract.

All other persons and organizations should apply to the:

U. S. DEPARTMENT OF COMMERCE
OFFICE OF TECHNICAL SERVICES
WASHINGTON 25, D. C.

FOREWORD

The enclosed paper is scheduled to appear in the volume Vistas in Information Handling and is being distributed in the present form in the interests of prompt reporting of work in progress.

Although some of the material covered was previously discussed in Sections III and IV of our first report,* this paper develops the linear method of association more simply and concisely. We have, for example, revised the mathematical development so that the linear transformations employed in the linear network approach to relevance-ranking are exhibited as matrices. This new formulation enables us to show that the relevance-ranking transformation can be represented as the product of three (intuitively meaningful) mappings in a fairly natural manner: an index term association mapping, a mapping from index term values to document values, and a document association mapping. In addition to mathematical simplifications, new material is also introduced on a hypothesis relating to the breakdown of "association" into a "synonymy" and a "contiguity" portion. These types of relationships are tentatively explored in terms of the matrix formulation. A portion of Section III of Report CACL-1 is included almost intact as an appendix to the paper; it is repeated in this report for completeness.

It is felt that the more compact presentation given in this paper will simplify and augment our earlier discussion of the linear network model, and the paper is distributed for this purpose. Since it treats only one of several topics under investigation, it should not be construed as a complete report of our activities under the contract.

ACKNOWLEDGMENT

The writers are indebted to Richard F. Meyer, also of Arthur D. Little, Inc., who was responsible for the initial formulation of a linear model for information retrieval.

---

* Arthur D. Little, Inc. Studies for the Design of an English Command and Control Language System, Report No. CACL-1, ESD Report No. TR-62-45.

LINEAR ASSOCIATIVE INFORMATION RETRIEVAL

Vincent E. Giuliano and Paul E. Jones

Arthur D. Little, Inc.

35 Acorn Park

Cambridge, Mass.

ABSTRACT

This paper is concerned with the recognition and exploitation of term associations for the retrieval of documents. A general theory of association and associative retrieval is presented; it is based on tne use of linear transformations, both for establishing associations among terms and for discriminating among documents. The design and behavior of a simple experimental device which realizes the theory is discussed.

LINEAR ASSOCIATIVE INFORMATION RETRIEVAL*

> "The real heart of the matter of selection, however, goes deeper than a lag in the adoption of mechanisms by libraries, or a lack of development of devices for their use. Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. When data of any sort are placed in storage, they are filed alphabetically or numerically, and information is found (when it is) by tracing it down from subclass to subclass. It can be in only one place, unless duplicates are used; one has to have rles as to which path will locate it, and the rules are cumbersome. Having found one item, moreover, one has to emerge from the system and re-enter on a new path.
>
> "The human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trials carried by the cells of the brain .... Man cannot hope to duplicate this mental process artificially, but he certainly ought to be able to learn from it.
>
> "As We May Think," by Vannevar Bush, from Atlantic Monthly, July 1945.

I. INTRODUCTION

This paper is concerned with the recognition and exploitation of term associations for the retrieval of documents. A general theory of association and associative retrieval is presented; it is based on the use of linear transformations, both for establishing associations between terms and for discriminating between documents. A simple experimental device which realizes the theory has been built, and examples of its operation are discussed in the Appendix.

A document retrieval system may be generally characterized in the following manner: a collection of d documents and a set of t index

* The work described herein has also been supported in part by the National Science Foundation

terms are presumed to be given, and it is assumed that each document has been indexed by the assignment to it of one or several applicable index terms.* Based upon such an indexing of the document collection, a retrieval system ideally functions so as to identify exactly those documents which are relevant to an inquiry consisting of a specification of one or more pertinent index terms.

To proceed more formally, an inquiry may be regarded as consisting of an assignment of positive "importance" values, not all necessarily equal, to some of those index terms which most directly characterize the matter of interest to the inquirer, and an assignment of value zero to all other index terms. It is convenient to consider the inquiry to be represented by a t dimensional column vector Q, in which each component q_i exhibits the value assigned to the ith index term by the inquirer.

Likewise, the response of a retrieval system to an inquiry will be regarded to be an assignment of nonnegative values for retrieval to all documents in the collection, where the values reflect relevance to the inquiry. This assignment defines a d dimensional response vector R, in which r_j exhibits the value assigned to a document j by the system in response to a given inquiry Q.

The retrieval process can therefore be viewed as a mathematical transformation from an inquiry vector Q to a response vector R, and there is no loss in generality in doing so.

---

* In practice, the index terms may be either descriptors assigned by manual indexing procedures, or keywords selected from the text by automatic means. This paper is not concerned with the relative merits of these two different philosophies of index term assignment, but rather is intended to suggest a method of retrieval which will yield improved results with either.

The relative magnitude of the values r_j define an ordering on the documents. In an effective retrieval system, the ordering of document values should reflect the actual ordering of relevance of the documents, and the documents should be retrieved in the decreasing order of this value.*

Associative Retrieval Methods

An "associative" retrieval system is one which attempts to take possible interconnections among index terms into account in performing the retrieval transformation. One objective of introducing association is to give, as a response, an ordering of documents ranked on a continuous scale of relevance rather than artificially grouped into two classes. Another objective is to free the requestor from a necessity to couch his inquiry in precisely the same terms employed by the indexer.

The role of term association within an associative retrieval scheme is roughly as follows. Each index term is regarded to bear some stronger or weaker measure of association with each other index term. A document is evaluated with respect to a given inquiry not only by considering the presence of index terms intersecting the document and inquiry, but also considering the presence of other terms in the document which may in turn be strongly associated with terms in the inquiry. Thus, for example, given an inquiry about the "production of "automobiles" a system may assign a high value to a document about the "manufacture" of "motor vehicles" provided that the corresponding terms are known to be highly associated.

---

* In most existing operational retrieval systems, of course, only two levels of value are recognized, "relevant" and "irrelevant. Such a boundary line is usually an artificial one, however, and hence is the source of many errors in retrieval; either "relevant" documents are not in fact relevant, or "irrelevant" documents are in fact relevant - a phenomenon well known to users of document retrieval systems.

A retrieval system embodying an automatic thesaurus thus qualifies as being "associative" in the strict sense stated above. In this paper however, we will be primarily concerned with the case in which the associations are based on formal statistical relationships present within a given document collection. A brief speculative discussion of possible linguistic interpretations which might be assigned to such formal associations will be given in Section IV.

Methods for performing retrieval transformations with the use of term associations have been discussed in the literature,[1,2,3,4] but, to our knowledge, the use of any but the most trivial linear transformations has never been proposed previously--a surprising fact since linear transformations are so well understood and since they have been applied so widely elsewhere. We shall explore the implications of assuming that the retrieval transformation is linear. Suppose that the linear transformation is represented by a d x t matrix Γ so that

$$R = \Gamma Q \tag{1}$$

It will be shown that Γ can be viewed as the product of three separate linear transformations, each of which has a meaning for retrieval. These will be represented by matrices Φ, Θ, and Λ so that

$$R = \Phi\Theta\Lambda Q \tag{2}$$

In this formula, Λ represents a t x t <u>index term association matrix</u>, so that the vector ΛQ represents a column vector of values of index terms after the effects of term association are taken into account; the value assigned to a term in ΛQ reflects how closely related that term is to the terms specified in the inquiry. Thus, for example, although in Q "production" may have value one hundred and "manufacture" value zero, in ΛQ "manufacture" will also have positive value, say eighty. The matrix Θ is a d x t matrix which attributes values to documents based

on the values of the terms. It is called a discriminant matrix, and the result $\Theta \cap Q$ is in the form of a document response vector. A final mapping, the d x d matrix ϕ, takes into account interactions (if any) among documents. Such interactions may arise, for example, when the paragraphs of an article or the chapters of a book are treated as distinct documents. Since the matrix ϕ performs a transformation which takes into account known associations among documents, it is called a document association matrix.

The overall retrieval transformation is thus viewed as a product of three linear transformations: an index term association transformation, a transformation from index term values to document values, and a document association transformation.

II. DETERMINING THE ASSOCIATION MAPPINGS

The linear associative transformations may be developed from at least three equivalent points of view:

1. Reasoning along probabilistic lines, in which term-term association is regarded to be a Markov process.
2. Reasoning based upon an electrical network analog.
3. Reasoning based upon the imposition of certain mathematical constraints on association and identification transformations, primarily consisting of certain assumptions of linearity and normalizability of transformation matrices.

All three approaches are ultimately equivalent in that they lead to the same set of mathematical formulas. The approaches differ in the interpretations they provide; each gives a different avenue of appeal to intuition. The third approach will be pursued in this Section, and the relationship to the electrical network analog approach will be outlined in Section III.

A. Case of the Index Term-Document Network

It will be assumed that every document i is connected to each index term j contained within that document by a bond of strength $C_{ij} > 0$, and that this strength can be determined by simple formal properties of the document and index term. This number might, for example, be given by the frequency of occurrence of the index term in the document, or it might possibly be determined with the aid of syntactic information.[4,5] The $d \times t$ matrix C_{ij} will be called the document-index term connection matrix for the corpus.

1. Conventional Term Retrieval

Almost all conventional (coordinate) retrieval systems rank documents for retrieval according to the value R obtained from the simple linear mapping:

$$R = C\,Q \tag{3}$$

In the usual case the values in both C and Q are restricted to 1 or 0, and it is evident that those documents having all the k index terms specified with value 1 in Q are ranked first, followed by those documents containing subsets of $k - 1$ index terms specified with value 1 in Q, etc.

It is clear, however, that the mapping (3) attributes nonzero values for retrieval only to documents having at least one term in common with the inquiry Q, and there is no provision for retrieval if synonymous or otherwise associated terms are used in the document and inquiry. Thus documents on the "production" of "automobiles" might be missed if an inquiry is phrased in terms of the "manufacture" of "motor vehicles."

2. Associative Retrieval

(a) Expansion in Powers of Index Term Connections

A formal approach to the association problem is to revise the mapping (3) to take into account index term interconnections as defined by the corpus itself. This may be accomplished by using powers of the term-term connection matrix $K = C^T C$. A typical element $k_{rs} = \sum_i C_{ir} C_{is}$ of this matrix gives a measure of interconnection between terms r and s via all documents that contain both of them. An element k_{rs} of K^2 gives a measure of interconnection between terms r and s via all pairs of documents such that one contains r, the other s, and both share one or more other index terms. Similarly, higher powers of K give measures of interconnection via longer and longer paths of documents and terms. Obviously, by taking a weighted sum of powers of K it is possible to obtain an association matrix which reflects the total effect of all paths of every length; indeed this is what will be done. For such a weighted sum of powers of K to be meaningful for retrieval purposes, however, it is first necessary to select the weights so that the strengths of association for shorter paths count more than those for longer paths. In fact, association strengths for longer and longer paths should approach zero, that is,

$$\lim_{n \longrightarrow \infty} \tilde{\tilde{K}}^n = 0 \tag{4}$$

if $\tilde{\tilde{K}}$ is a properly normalized (weighted) term-term connection matrix.

(b) Selection of a Normalization

Since all elements of K are non-negative, a sufficient condition for convergence (4) is that all of the rows of $\tilde{\tilde{K}}$ sum to less than unity. This will hold, in turn, if

$$\tilde{K} = C^T \tilde{C} \tag{5}$$

$$\tilde{\tilde{K}} = \lambda \tilde{K} = \lambda C^T \tilde{C} \tag{6}$$

where the row sums of $\tilde{C}^T$ and $\tilde{C}$ are normalized to unity, and where λ is a t by t diagonal matrix with all $0 \leq \lambda_{ii} < 1$. Specifically, we take

$$\tilde{C} = \sigma C \text{ and } \tilde{C}^T = \rho C^T \tag{7}$$

where σ and ρ are diagonal matrices given by

$$\sigma_{ii} = \frac{1}{\sum_j C_{ij}} \quad \text{and} \quad \rho_{jj} = \frac{1}{\sum_i C_{ij}} \tag{8}$$

(note $\tilde{C}^T \neq \tilde{C}^T$)

A normalization constant $\lambda_{ii} < 1$ is presumed to be assigned to each index term so that, very roughly speaking, $\frac{1}{\lambda_{ii}} - 1$ determines the cost of associating from one document to another through index term i. The intuitive meaning of this normalization will become clearer to the reader as he proceeds through the discussion of the remainder of this Section, Section III, and the Appendix.

(c) Assumptions of Linearity

Now we are prepared to obtain the actual form of the linear associative retrieval mapping, which follows from two explicit assumptions of linearity:

a. The value of a document is a linear function of the values of the index terms contained in it, where the coefficients of the function are given by $\tilde{C}$. That is, if W is the vector of index term values,

$$R = \tilde{C} W \tag{9}$$

b. The value of an index term is given by taking the sum of its original value assigned by the inquiry Q and a linear function λC^T of the values of the documents containing it, specifically:

$$W = \lambda \tilde{C}^T R + Q \tag{10}$$

The desired retrieval mapping follows directly from (9) and (10):

$$\boxed{R = \tilde{C} \left[I - \lambda \tilde{K} \right]^{-1} Q = \tilde{C} \left[I - \lambda C^T \tilde{C} \right]^{-1} Q} \tag{11}$$

which can, given (4), be written as a convergent series:

$$\boxed{R = \tilde{C} \left[I + (\lambda \tilde{K}) + (\lambda \tilde{K})^2 + (\lambda \tilde{K})^3 + \ldots \right] Q} \tag{12}$$

Equation (12) is the desired mapping which takes into account the weighted and summed effect of association paths of all lengths. It is obviously a generalization of the conventional retrieval mapping (3). It may be noted first of all that this is of the form (2) where $\tilde{C} = \Theta$ is the discriminant matrix, $\left[I - \lambda \tilde{K} \right]^{-1} = \Lambda$ is the index term association matrix* and since there are no direct document interconnections, $\Phi = I$, the identity mapping. It may be noted next that rapidity of convergence is determined by λ, and that for $\lambda = 0$, the conventional

* Matrices of this form have been applied to the study of indirect interactions among sectors of the economy by Leontief and others,[6] and are often referred to in the literature as "Leontief Matrices".

term retrieval mapping (3) is obtained. By regulating the values of λ, association can be either "free" (for λ's near 1) or "narrow" (for λ's near 0). The effectiveness of retrieval using this method will be discussed in Section V.

B. Case with Interdocument Linkages and Inter-Index Term Linkages

The discussion in the previous Subsection presumed the absence of direct linkages between index terms and other index terms, as well as the absence of direct linkages between documents and other documents. Introduction of such direct linkages may prove useful, however, for two principal reasons: First of all, if documents are themselves interrelated such as by being chapters, sections, or even paragraphs within a given book, it may be desirable to reflect these relationships with inter-document links. Information present in citations might also be used to generate such inter-document links. In the extreme case, when individual documents are sentences in a stream of writing or speech, strong intersentence linkages can provide in part for the effect of antecedence.

Secondly, it could conceivably be desirable to establish a priori linkages between synonyms. Such linkages would provide a direct means of introducing "semantic" information not present in the corpus itself.*

The remainder of this Section is therefore concerned with generalizing the mathematical treatment of Part A to include the situation when direct document-document links and direct term-term links are permitted. In

---

* Although the presentation is extended in this Subsection to accommodate the case when a priori links between index terms are presumed given, there is reason to believe that such "thesaurus" entries will in fact be unneeded; such has been the general experience of other workers on associative indexing schemes.[1,2,7] The writers conjecture, in fact, that the linear transformations to be developed in this Subsection might be used to generate a "thesaurus" valid for a given corpus directly from the C matrix for that corpus.

this more general case the connection matrix is square of dimension $d + t$ partitioned as follows:

$$G = \begin{pmatrix} A_{dd} & C_{dt} \\ B_{td} & D_{tt} \end{pmatrix} \tag{13}$$

In this representation, C is the document-term connection matrix employed before, B is its transpose, A the document-document connection matrix, and D the term-term connection matrix. The normalization is obtained by multiplying on the left by a diagonal matrix, analogous to (8) giving:

$$\begin{pmatrix} \sigma_{dd} & 0 \\ 0 & \rho_{tt} \end{pmatrix} G = \begin{pmatrix} \tilde{A} & \tilde{C} \\ \tilde{B} & \tilde{D} \end{pmatrix} \tag{14}$$

where

$$\sigma_{ii} = \frac{\lambda_{ii}}{\sum_{j=1}^{d} A_{ij} + \sum_{j=d+1}^{t+d} C_{ij}} \quad \text{and} \quad \rho_{ii} = \frac{\lambda_{ii}}{\sum_{j=1}^{d} B_{ij} + \sum_{j=d+1}^{t+d} D_{ij}}$$

Again, in this more general case we still require the solution to satisfy linear constraints analogous to (9) and (10)

$$\begin{pmatrix} R \\ W \end{pmatrix} = \begin{pmatrix} \tilde{A} & \tilde{C} \\ \tilde{B} & \tilde{D} \end{pmatrix} \begin{pmatrix} R \\ W \end{pmatrix} + \begin{pmatrix} 0 \\ Q \end{pmatrix} \tag{15}$$

and therefore:

$$\begin{pmatrix} R \\ W \end{pmatrix} = \begin{pmatrix} I - \tilde{A} & -\tilde{C} \\ -\tilde{B} & I - \tilde{D} \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ Q \end{pmatrix} \tag{16}$$

This can be solved to give the generalized association formulas:

$$\boxed{R = (I - \tilde{A})^{-1} \tilde{C} (I - \tilde{D})^{-1} \left[I - \tilde{B} (I - \tilde{A})^{-1} \tilde{C} (I - \tilde{D})^{-1} \right]^{-1} Q} \tag{17a}$$

and

$$\boxed{W = (I - \tilde{D})^{-1} \left[I - \tilde{B} (I - \tilde{A})^{-1} \tilde{C} (I - \tilde{D})^{-1} \right]^{-1} Q} \tag{17b}$$

Equation (17) may be recognized to be in the form of equation (2). The factor $\Lambda = (I - \tilde{D})^{-1} \left[I - \tilde{B} (I - \tilde{A})^{-1} \tilde{C} (I - \tilde{D})^{-1} \right]^{-1}$ gives the generalized term association mapping, taking into account term-term connections, term-document connections, and document-document connections. The factor $C = \Theta$ gives the discriminant mapping from term values to document values. Finally, the $(I - \tilde{A})^{-1} = \Phi$ factor gives a document association mapping, which is performed on the document values. Obviously, (17) is a generalization of (11) which is a generalization of the conventional retrieval mapping (3).

In some applications it may be desirable to decouple the effects of the document association mapping and the discriminant mapping from those of the term association mapping. This can be accomplished by choosing the c_{ij} and b_{ij} values to be small compared to the a_{ij} and d_{ij} values. In this case (17) essentially reduces to

$$R = (I - \tilde{A})^{-1} \tilde{C} (I - \tilde{D})^{-1} Q \qquad (18)$$

indicating that for practical purposes Φ, Θ, and Λ can be chosen independently.

Automatic Thesaurus Generation

The transformation (17b) can be used to generate a thesaurus-like listing valid for the given document collection. Suppose that no a priori synonym linkages are given (i.e., $D = 0$), and that Q_z is a unit vector assigning value only to index term z. Then (17b) becomes

$$W_z = \left[I - \tilde{B} (I - \tilde{A})^{-1} \tilde{C} \right]^{-1} Q_z$$

where the values in W_z rank every index term according to its degree of association with index term z. By listing the few topmost-ranked terms in W_z for each term z in turn, a "thesaurus" listing can be obtained completely automatically. The validity or usefulness of such a listing must of course be established by experiment.

III. THE ELECTRICAL NETWORK ANALOG

In principle at least, the general linear retrieval transformation (17) can always be represented and solved by means of an electrical network analog. To envisage the network which is the analog of (17), imagine two sets of electrical binding posts, one post for each document and one post for each index term. Suppose now that a resistor with conductance g_{ij} is soldered between every pair of binding posts i and j, where g_{ij} is the connection matrix element in (13). Also suppose that a "leak" resistor with conductance C_{oj} is soldered between each index term post j and some common return post 0.

To pose an inquiry Q with components q_f, current is to be injected into the index term binding posts in the quantity:

$$J_{\substack{\text{injected}\\ \text{into f}}} = q_f \left(C_{of} + \sum_{i=1}^{d} B_{if} + \sum_{i=d+1}^{d+t} D_{if} \right)$$

Then we claim that the R values will be the voltages appearing in the document binding posts. That this is so can be seen by writing the equations which govern the behavior of the network.

By conservation of current at any document node p we have:

$$\sum_{j=1}^{d} J_{pj} + \sum_{j=d+1}^{t+d} J_{pj} = 0 \qquad (19)$$

Now let r_p be the voltage on a document node p, w_j be that on an index term node j. We then have, writing Ohm's law out using the notation of (13):

$$\sum_{j=1}^{d} A_{pj} (r_j - r_p) + \sum_{j=d+1}^{t+d} C_{pj} (w_j - r_p) = 0 \qquad (20)$$

giving:

$$r_p = \frac{\sum_{j=1}^{d} A_{pj}\, r_j + \sum_{j=d+1}^{t+d} C_{pj}\, w_j}{\sum_{j=1}^{d} A_{pj} + \sum_{j=d+1}^{t+d} C_{pj}} \qquad (21)$$

Likewise, at any term binding post, f, we have:

$$\sum_{i=1}^{d} J_{if} + \sum_{i=d+1}^{t+d} J_{if} + J_{injected} = 0 \qquad (22)$$

Applying Ohm's law,

$$C_{of}\, w_f + \sum_{i=1}^{d} B_{if}\, (r_i - w_f) + \sum_{i=d+1}^{t+d} D_{if}\, (w_i - w_f) + J_{injected} = 0 \qquad (23)$$

giving:

$$w_f = \frac{\sum_{i=1}^{d} B_{if}\, r_i + \sum_{i=d+1}^{t+d} D_{if}\, w_i}{C_{of} + \sum_{i=1}^{d} B_{if} + \sum_{i=d+1}^{t+d} D_{if}} + q_f \qquad (24)$$

It remains only to note that equations (21) and (24) are formally the same as those given in (15).

The constants λ_{jj} and C_{oj} are related in that:

$$\lambda_{jj} = \frac{\sum_{j=1}^{d} B_{ij} + \sum_{j=d+1}^{t+d} D_{ij}}{C_{oj} + \sum_{j=1}^{d} B_{ij} + \sum_{j=d+1}^{t+d} D_{ij}} \qquad (25)$$

Obviously, by regulation of the "leak" conductance C_{oj} any value of λ_{jj} can be obtained between 1 and 0. The physical network therefore always has a solution as long as there is at least one path from every

terminal into which current might be injected and ground. Therefore, as seen previously, a sufficient condition for equations (17) and (11) to have solutions is that all $\lambda_{jj} < 1$. Finally, a linear associative retrieval transformation (17) can in principle always be realized by means of an electrical network.

IV. THE NATURE OF ASSOCIATIONS DERIVED FROM TEXT

A mathematical apparatus has been proposed which, among other things, is capable of generating measures of association between index terms present in a given body of text. It is interesting to speculate of the nature of some of the linguistic factors which these association measures may reflect.

Hopefully, one such factor will be that of semantic overlap and partial synonymy, such as is exhibited by the pair of terms "production" and "manufacture" or by the pair "chair" and "seat." The notion of synonymy association has strong intuitive appeal, and associations of this kind tend to be relatively permanent features of the language.

A second association-producing factor also exists, however; it relates not to synonymy but rather to real-world relationships between the objects or actions designated by the index terms. Associations due to this factor, for example, are exhibited in the relationships among "atom," "warhead," "bomb," missile, and in the relationship between "satellite" and "Cape Canaveral." Psychologists sometimes refer to these associations as being due to "contiguity." They tend to be impermanent and to be strongly conditioned by the nature of one's experience. The association between "satellite" and "Cape Canaveral" could, for example, disappear once again if satellites were to be launched primarily from a new location. And while the general public today might agree that "satellite" and "Cape Canaveral" are more highly associated than "satellite"

and "telemetry," a population of electronic engineers might have the opposite point of view.

There is little doubt that information about both synonymy association and contiguity association could be exploited within the context of a document retrieval system, if this information could be made readily available. As yet, however, there is relatively little experimental evidence to indicate the extent to which the associations generated by the linear association process (or any other association process, for that matter) reflect either contiguity or synonymy. Nonetheless, the writers feel that it is appropriate to devote a few paragraphs to conjectures which seem to indicate that the linear process can be used to generate associations either due to the contiguity and synonymy factors in combination, or due to either of these factors separately.

These conjectures are based on interpretation of the coefficients in the right hand side of the power series expansion of the index term association matrix, from (12):

$$(I - \lambda\tilde{K})^{-1} = I + \lambda\tilde{K} + (\lambda\tilde{K})^2 + (\lambda\tilde{K})^3 + \ldots \qquad (26)$$

The leading coefficient I generates the identity association; the fact that any index term is most highly associated with itself is of course a truism from both the viewpoints of synonymy and of contiguity. The second coefficient, λK, generates a measure of association between a pair of terms which depends on the number of times they have co-occurred in the given body of documents. We speculate that this coefficient primarily reflects contiguity association, since a pair of index terms found in a certain document will in general describe things that have to do with one another in fact; indeed, what they have to do with one another is often the subject of the document. Depending on the methods and conventions used for indexing, of course, partially synonymous

index term may also be used to characterize a given document, but this effect can still reasonably be expected to be overshadowed by that of contiguity. A first conjecture, then, is that the $\Lambda\tilde{K}$ term primarily reflects association due to contiguity. Considering association due to this contiguity factor $\Lambda\tilde{K}$ alone, any given index term will bear a stronger or weaker measure of $\Lambda\tilde{K}$-association with every other index term. For any given index term, the set of $\Lambda\tilde{K}$ associations it bears with all other index terms will be called its contiguity profile.

It is apparent that the third coefficient in the series expansion $(\Lambda\tilde{K})^2$ generates a measure of association between index terms which depends only on the similarity of their contiguity profiles. That is, $(\Lambda\tilde{K})^2_{ij}$ will be large if and only if index terms i and j have similar contiguity profiles. At this point, it can be conjectured plausibly that synonymous index terms can be identified and associated by the similarity of their contiguity profiles. Indeed, arguments in favor of this second conjecture have frequently been advanced by structural linguists. Presuming its validity, it follows that the $(\Lambda\tilde{K})^2$ coefficient gives a measure of association due primarily to synonymy.

Moreover, if both of the above conjectures are valid, it follows that all of the odd powers of $\Lambda\tilde{K}$ in the series expansion primarily represent association due to contiguity, all of the even powers of $\Lambda\tilde{K}$ represent association primarily due to synonymy. The association matrix we have been dealing with so far (26) therefore reflects association due to both contiguity and synonymy factors in combination. There is nothing, however, to prevent use of other transformations which represent these factors separately. In particular, the association matrix

$$\Lambda K \left[I - (\Lambda K)^2 \right]^{-1} = \Lambda K + (\Lambda K)^3 + (\Lambda K)^5 + \ldots \qquad (27)$$

contains only odd powers, and according to the above conjecture primarily represents contiguity association, the association matrix

$$\left[I - (\lambda K)^2\right]^{-1} = I + (\lambda K)^2 + (\lambda K)^4 + (\lambda K)^6 + \ldots \quad (28)$$

contains only even powers, and primarily represents synonymy association. A subsequent paper will deal with these points further, and will describe experimental evidence relating to these conjectures.

If associations are based on a finite body of text, there is an important third association-producing factor in operation which depends on the statistics of the usage of the terms in indexing the collection. Suppose, for example, that in a certain small set of documents, "paper" and "tractor" happen to be used as index terms in characterizing exactly the same subset of documents. The inadequacy of the sample of documents thus leads to the conclusion that "paper" and "tractor" are strongly associated. This is, of course, what we desire in a retrieval context. The two terms are wholly redundant for retrieval since the use of either term in an inquiry will cause the retrieval of exactly the same documents. An analogous situation arises when the overlap is only partial. But this third factor points to a problem--even if associations derived from text can be expected to yield valid measures of contiguity and synonymy relationships, this will be so only when the body of text is sufficiently large to give statistically meaningful results.

V. EXPERIMENTAL WORK

A. The ACORN Devices

Solution of equation (17) by digital techniques involves multiplication and inversion of very large matrices--tedious processes even when a very high speed digital computer is available. Of course, computational short cuts exist when the matrices are very sparse as these matrices are, but nonetheless a very large amount of computing is still required if the processing is to be done digitally.[6] For this

reason we have begun to investigate the practical use of analog electrical networks which solve the linear equations directly.

Two simple experimental devices for linear association have been built and are undergoing testing--we have called them ACORN-I and ACORN-II, standing for "Associative Content Retrieval Network." Although the ACORNS are both basically small scale demonstration devices, some interesting preliminary experiments are being done with them.

ACORN-I is shown in Fig. 1; it presently accommodates a total of 82 sentences (which represent documents) and index terms. Each index term and each sentence is represented by a terminal; and the terminals are interconnected by resistors as shown. The light-colored wires terminate with alligator clips, and are attached to the terminals for the key words in the inquiry. The relative voltages on these wires are controlled by the potentiometer knobs shown. As the overall voltage is raised by turning the large knob, current injected into the network is increased and the neon bulbs connected to the sentence terminals light up in the order of "relevance" determined by the network. Some examples of the behavior of ACORN-I will be discussed in the Appendix.

ACORN-II shown in Fig. 2, is presently wired for the associative "retrieval" of libraries in the area of political science; it accommodates 110 libraries and 48 topics. The topics are represented by the terminals on the inner ring, the libraries by the terminals and neon bulbs on the outer ring. The complete circuits of both of these ACORN devices, except for the normalization resistors, are shown on the front panels; the only electronic components required are resistors and neon bulbs.

B. Preliminary Retrieval Experiments

In the theoretical discussion, we have given only a formal characterization of the ordering induced on documents in response to a question. In general, the problem of measuring the effectiveness of such an ordering for purposes of retrieval is an extremely difficult one, and one which merits considerable further study. Likewise, a great deal still remains to be learned about the behavior of the linear networks under different conditions of normalization and inquiry configuration. Nonetheless, it is possible to characterize the behavior of the ACORN devices under the specific wiring configuration employed in their construction. This behavior is illustrated in the Appendix by means of some exercises run on ACORN-I, and is summarized in the following paragraph. Although the linear associative technique has not yet been tested on a large scale, there seem to be excellent reasons for assuming that, through proper selection of normalization, the behavior patterns of ACORN-I can just as well be realized for a much larger number of documents and larger number of index terms.

In summary of the Appendix, ACORN-I behaves as if the retrieval ordering were produced as a result of the interaction of four main factors which, stated in simplified form in order of decreasing importance, are:

1. The number of key terms shared by inquiry and sentence; the sentences containing the most key terms specified in the inquiry will in general be listed first.

2. The frequency of use of key terms. If several key terms are specified in an inquiry, sentences containing a given number of the more rarely used of these key terms will tend to be ranked above those containing the same number of frequently used key terms.

3. The number of extraneous index terms in a sentence. A sentence containing a certain number of key terms specified in the inquiry and

no other key terms will tend to be ranked before one containing the same key terms plus extraneous (nonassociated) ones.

4. The degree of indirect association between index terms in a sentence and key terms in an inquiry. Other factors being equal, sentences containing index terms associated with those in the inquiry are ranked above others.

The patterns according to which these factors appear to interact are discussed further in the Appendix; they are fairly complex, but on the whole quite pleasing. It would clearly be desirable to have a large scale document retrieval system behave according to these patterns, and the writers are therefore planning tests of the method on a much larger scale.

C. Additional Features of Interest

Three further points of interest regarding the ACORN networks will be mentioned in this paper. The first point is that an ACORN analog network can be used to perform several functions which are in practice very difficult to do with a conventional computerized document retrieval system; for example: (1) Retrieving a set of documents most closely related to one or more given documents. (2) Retrieving a set of index terms ('thesaurus generation'). (3) Retrieving a set of documents and index terms most closely related to a given combination of index terms and documents, etc.

The retrieval of a set of index terms, for example, may be accomplished by applying currents to the terminals for the given index terms, by reading voltages on the terminals for all other index terms, and by selecting those terms with highest voltages on their terminals.

The second point of interest is that there is no need to confine an ACORN to two levels of elements; a three-level device, for example, could recognize index terms, documents and scientists. Linear association and discrimination, as a matter of fact, can be extended to any number of levels. The third and final point relates to the remarkable insensitivity of the association mapping to variations in document-term connection strengths. Pulling out or cutting a few randomly selected wires in an ACORN generally has a surprisingly small effect. As a matter of fact, neither of the ACORN machines has had its connections verified or 'debugged' after its initial wiring; both worked when they were first plugged in. It was not until ACORN-II had been used for several public demonstrations, in fact, that it was discovered, quite by accident, that three of the wires were put in incorrectly--the effect of the errors only showed up for highly special request combinations, and then only in disturbing the ordering of the third, fourth, and subsequent items retrieved! This insensitivity is of course explainable in terms of the multiplicity of indirect and redundant association paths which remain intact when a direct path is severed. Obviously, this insensitivity is of value because it enables use of wide-tolerance components in constructing ACORN devices. It also suggests that the retrieval process can indeed be made insensitive to minor variations in indexing--one of the practical objectives which has motivated the work described in this paper.

REFERENCES

1. Maron, M. E., and Kuhns, J. L. "On Relevance, Probabilistic Indexing and Information Retrieval," Journal of the ACM, Vol. 7, 1960, pp. 216-244.

2. Stiles, H. E., "The Association Factor in Information Retrieval," Journal of the ACM, Vol. 8, 1961, pp. 271-279.

3. Doyle, L. B., "Semantic Road Maps for Literature Searchers," Journal of the ACM, Vol. 8, 1961, pp. 553-578.

4. Studies for the Design of an English Command and Control Language System (Giuliano, Clapp, Jones, Kuno, Meyer, Oettinger, Rubenstein, Sherry), Report ESD-TR-62-45. OAL, ESD, AFSC, U.S. Air Force, 1962.

5. Salton, G., "Some Experiments in the Generation of Word and Document Associations," (draft) Harvard Computation Laboratory.

6. Clark, P. G., Interindustry Economics, Wiley, 1959.

7. Borko, H., Discussion of a Proposed Study of Associations Derived from Text, System Development Corporation, SDC-FN-6081, December, 1961.

APPENDIX

RETRIEVAL BEHAVIOR OF ACORN-I

ACORN-I is currently wired for the associative retrieval of 42 sentences using 42 index terms. For the purpose of giving examples, the first 26 sentences and the first 24 index terms are listed in Table IA; the portion of the retrieval network employed is diagrammed in Table II. To avoid confusion we will refer to an index term which appears in an inquiry as a "key term."

In ACORN-II, all normalization conductances (C_{oj} in equation (25)) were chosen to be equal (22K), and all connection conductances (B_{ij} in equation (25)) were chosen to be equal to one of two values, (4.4K or 2.2K as shown in Table II.) All normalization resistors for index terms feed into a single potentiometer which provides a controllable "degree of association," since in effect it enables partial control of λ in equation (11). With this knob in its minimum position, λ is small and association is relatively "narrow." With this knob in its maximum position λ is large, and association as relatively "free." In the latter case, associated terms often count more than key items in determining relevance of a sentence. To illustrate the operation of the other factors besides association which determine the ordering, all of the example exercises discussed were made with the association control set to 'narrow."

1. Number of Key Terms Shared by the Question and Sentence

When association is set to "narrow" on ACORN-I, the most important factor in determining the ranking of a sentence in response to a given question appears to be the number of key terms shared by the question and sentence. Given a question containing m key terms, the sentences listed first will be those, if any, containing all m key terms; next any sentences containing m - 1 of the key terms will be listed, then

sentences containing m - 2 of the m key terms, etc., until finally, at the bottom of the list, sentences containing none of the key terms. If association is kept narrow, this first factor appears to exercise the most influence in determining the orderings; the other factors serve primarily to determine the ordering within a subset of sentences having a given number of the key terms.

Example (1)

What Do You Know About (W. D. Y. K. A.) Surveillance, Tactical Missions, and Continental Defense?

Experimentally observed answers, in order of decreasing relevance given by ACORN-I:

Sentences (20), (16), (17), (26). (24), etc.

Explanation of ordering:

Sentence (20) is the only one containing all three specified key terms.

Sentences (16) and (17) each contain two of the specified key terms.

Of these, sentence (16) contains only one other index term, and sentence (17) contains two other index terms.

All the lower-ordered sentences contain only one of the specified key terms.

2. The Number of Extraneous Index Terms in a Sentence

Given narrow association, a second factor which appears to be of importance in determining the ordering in ACORN-I, is the total number of index terms in any sentence as compared to the number of key terms; this is illustrated in the example just given. Given a subset of sentences, each of which contains exactly the same j key terms from the question, the ones having the fewest additional index terms will tend to be ranked first. That is, those sentences having exactly the j key terms and no others will be listed first, then those having the j key terms, plus one other index term will be listed next, then those

having the j key terms, plus two other index terms will be listed next, etc. The same factor tends to operate for sentences having different subsets of j key terms; a short answer containing a given number of key terms is treated by the network as preferable to a longer answer containing those key terms plus other index terms.

Example (2)

W. D. Y. K. A. Missile Interception?

Experimentally observed answers, in order of decreasing relevance as given by ACORN-I:

Sentences (5), (6), (25), etc.

Explanation of ordering:

All three sentences contain the specified key term.

Sentence (5) contains no other index terms.

Sentence (6) contains two other index terms.

Sentence (25) contains three other index terms.

3. The Frequency of Use of Key Terms

Again assuming narrow association, a third factor which tends to influence the order of listing of sentences in ACORN-I is the frequency of use of the specified key terms within the corpus as a whole. Given a set of sentences each of which contains exactly j out of m key terms, those containing the least frequently used key terms will tend to be listed first. For example, given a set of sentences each of which contains one out of several key terms in the question and no oth r index terms, those sentences containing the rarely used key terms will be listed before those containing the more frequently used key terms. This third factor generally takes precedence over the second factor in determining the ordering, but it appears to be considerably less significant than the first factor.

Example (3)

W. D. Y. K. A. Air Force and SAC?

Experimentally observed answers, in order of decreasing relevance as given by ACORN-I:

Sentences (22), (18), (1), (9), (8), (13), (10), (7), etc.

Explanation of ordering:

Since no sentence in the corpus contains both of the specified key terms, those listed contain only one of the specified key terms.

The two sentences containing the more rarely used specified key term, SAC, are listed first; of these (22) contains only two other index terms and is listed first, (18) contains four other index terms, and is listed second.

All the remaining listed sentences contain Air Force, a frequently used index term.

4. Associations Among Terms

The fourth and most subtle factor operating to influence the ordering in ACORN-I is the association itself; this factor relates to how closely index terms in a sentence are associated with key terms via other sentences. Consider, for example, the subset of sentences which have no index terms in common with the question. Even with the association kept "narrow," these sentences will still be ordered with respect to one another, their ranking being dependent on how closely the index terms they contain are associated with terms appearing in the question. Those sentences containing the most highly associated terms will be listed first.

Example (4)

W. D. Y. K. A. Submarines?

Experimentally observed answers, in order of decreasing relevance as given by ACORN-I:

Sentences (23), (24), (11), (13), (10), (12), etc.

Explanation of ordering:

Sentences (23) and (24) are the only ones containing the key term submarines.

(23) is listed first because it contains fewer other index terms.

Sentences (11), (13), (10), and (12) all contain the index term Navy which is linked to submarines via both sentences (23) and (24).

The fourth factor operates in interaction with the first, second and third factors, its relative importance depending on the setting of the degree of association knob." With association kept "narrow" the short indirect paths will have dominant influence in determining association. Its influence when the association is set to narrow" appears to be relatively weak, but still useful when the other factors do not pertain. The interaction of all four factors can be seen in the following example:

Example (5)

W. D. Y. K. A. Air Force and Military Establishment?

Experimentally observed answers, in order of decreasing relevance as given by ACORN-I:

Sentences (1), (10), (11), (13), (12), (2), (6), (25), (23), (9), (16), (7), (17), etc.

Explanation of ordering

Sentence (1) contains both specified key terms and no other index terms. Sentence (10) contains both specified key terms but two other index terms as well.

All remaining sentences contain only one of the two specified key terms.

Sentence (11) is the only one of these containing military establishment, by far the most rarely used of the two specified key terms.

All remaining sentences listed above contain Air Force.

Of these (13) and (12) have the strongest indirect links to military establishment via the index term Navy which appears also in (11). Sentence (13) is listed before (12) because the former contains a total of three index terms; the latter contains four.

The desirability of the demonstrated interaction of these four factors in determining the ordering is clear. The first factor guarantees that the very top-most portion of the list of sentences will be precisely that set of sentences which would be delivered by a conventional term-superposition retrieval logic, i.e., those sentences containing all of the specified key terms. Obviously, in the absence of additional information, it is desirable to have listed next those sentences which have one less than the full complement of key terms, etc.

The second factor gives, assuming all else is equal, precedence to those sentences having fewer possible irrelevant index terms. This will result in the briefer sentences having the desired key terms being listed first; again, arguments have often been advanced in favor of such a retrieval policy.

The third factor gives heavier weight to the less frequently used index terms and is therefore desirable because these terms convey more information for retrieval purposes. For example, the sample corpus contains the index term Air Force very frequently, being about the Air Force. Accordingly, when the term Air Force is used in an inquiry, it is probably of considerably less importance than the other key terms in telling what the inquiry is about.

The fourth factor brings word associations into the retrieval picture, and is of great interest for it offers hope of freeing the inquirer from rigid and constrained use of index terms. This can perhaps best be illustrated by means of an example.

Example (6)

Suppose that an inquirer knows that the sentences he is looking for have most to do with missile interception:

W. D. Y. K. A. Missile Interception?

Experimentally observed answers in order of decreasing relevance as given by ACORN-I.

Sentences (5), (6), (25), (2), (12), etc.

Now let us suppose that another inquirer has the same interests as the first inquirer, but does not know enough to use the term missile interception (such might be the case if there were thousands of recognized index terms). Suppose, instead, that he asks a question with two key terms, Overland and Air Defense.

W. D. Y. K. A. Overland Air Defense?

Experimentally observed answers, in order of decreasing relevance as given by ACORN-I.

Sentences (2), (6), (25), (12), (5), etc.

Note that the same sentences are listed, in slightly different order.

It must be remarked that the indirect word associations do not always work out so neatly as in the last example, but this is to be expected since the sample corpus is far too small to represent the language.

TABLE IA

PORTION OF A DEMONSTRATION CORPUS

FOR ACORN-I

1. The Air Force (1) is a basic component of the military establishment (3).
2. National security (17) in our epoch requires extensive facilities for air defense (5), and the role of our Air Force (1) is to provide these.
3. The Air Force (1) maintains tight control over its nuclear warheads (2) by means of a variety of intricate procedures.
4. We have agreed to furnish nuclear warheads (2) to certain NATO (18) member nations, but only under tightly specified conditions.
5. The problem of missile interception (4) is exceedingly difficult, and it is possible that no effective solution may be found for years.
6. An extremely important problem of air defense (5) currently faced by the Air Force (1) is missile interception (4).
7. The Air Force (1) must exercise tighter operational control over ICBM's (12) and transonic bombers (22) than over conventional aircraft (6).
8. One of the most serious problems the Air Force (1) faces in developing satellite missiles (7) is the discovery of effective but yet fail-safe means for command and control (8) of them.
9. The Air Force (1) is developing integrated systems for command and control (8) of ICBM's (12) and transonic bombers (22), as well as other types of aircraft (6).
10. The three central components of the military establishment (3) are the Air Force (1), the Army (11), and the Navy (9).

11. The branches of the military establishment (3) primarily responsible for intermediate range missiles (10) are the Air Force (1) and the Navy (9).
12. Over-water air defense (19, 5) is primarily a joint responsibility of the Navy (9) and the Air Force (1).
13. Strenuous efforts are being made to facilitate improvement of communication between the Air Force (1), the Army (11), and the Navy (9).
14. Programs are being pursued for the further development and testing of compact yet powerful nuclear warheads (2) for ICBM's (12).
15. Intelligence reports indicate that the Soviets (20) may be behind in the development of compact nuclear warheads (2) for second generation ICBM's (12).
16. Some feel that the importance of aerial surveillance (13) and of tactical missions (14) is too often neglected in the Air Force (1) today.
17. Functions of the Air Force (1) include aerial surveillance (13), the maintenance of a strategic deterrence (15), and an ability to perform tactical missions (14).
18. The continuing development of our capability for strategic deterrence (15) will eventually require that the aircraft (6) of the SAC (16) be replaced by ICBM's (12), and perhaps eventually by satellite missiles (7).
19. Mainstays of the Air Force (1) capability for strategic deterrence (15) are our transonic bombers (22) and, increasingly, our ICBM's (12).
20. Effective continental defense (21) requires that our force for strategic deterrence (15) be strongly supported with capabilities for maintaining effective surveillance (13) and for executing tactical missions (14).

21. Aircraft (6) are still effective for providing strategic deterrence (15); the new transonic bombers (22), in particular, will be useful for several years to come.

22. The SAC (16) has an operational command and control (8) system used for aircraft (6) guidance.

23. The Navy (9) is authorized to carry both aircraft (6) and intermediate range missiles (10) on submarines (24).

24. A primary surveillance (13) role played by Navy (9) aircraft (6) is the detection of submarines (24).

25. Over-land (23) missile interception (4) is primarily the joint responsibility of the Army (11) and the Air Force (1).

26. Nuclear warheads (2) produced for applications of strategic deterrence (15) may be in the 500 kiloton and larger range, while those made for tactical missions (14) are generally smaller.

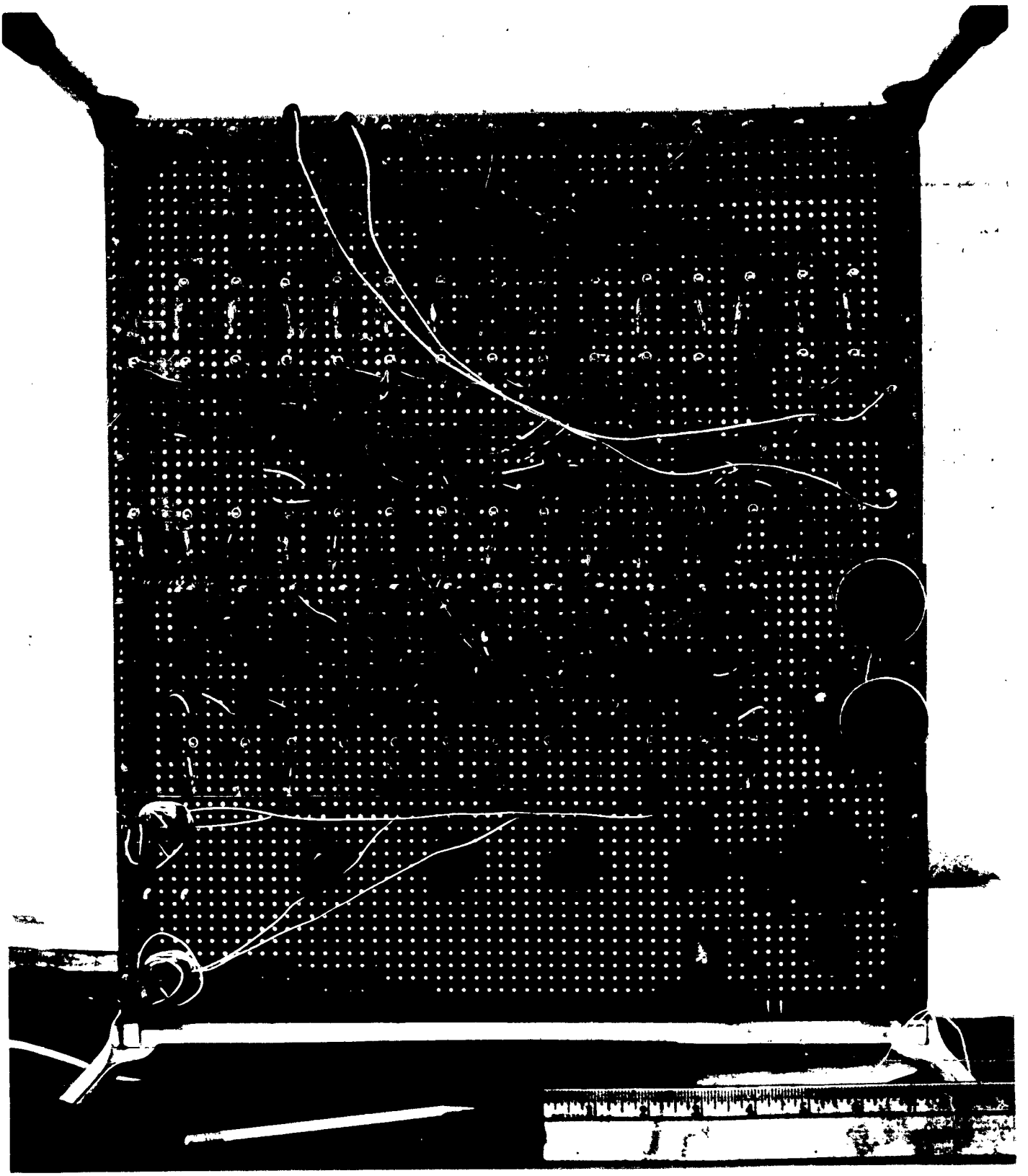
(The above sentences are intended for demonstration purposes only; they are not necessarily accurate in factual content and do not represent statements of opinion.)

TABLE IB

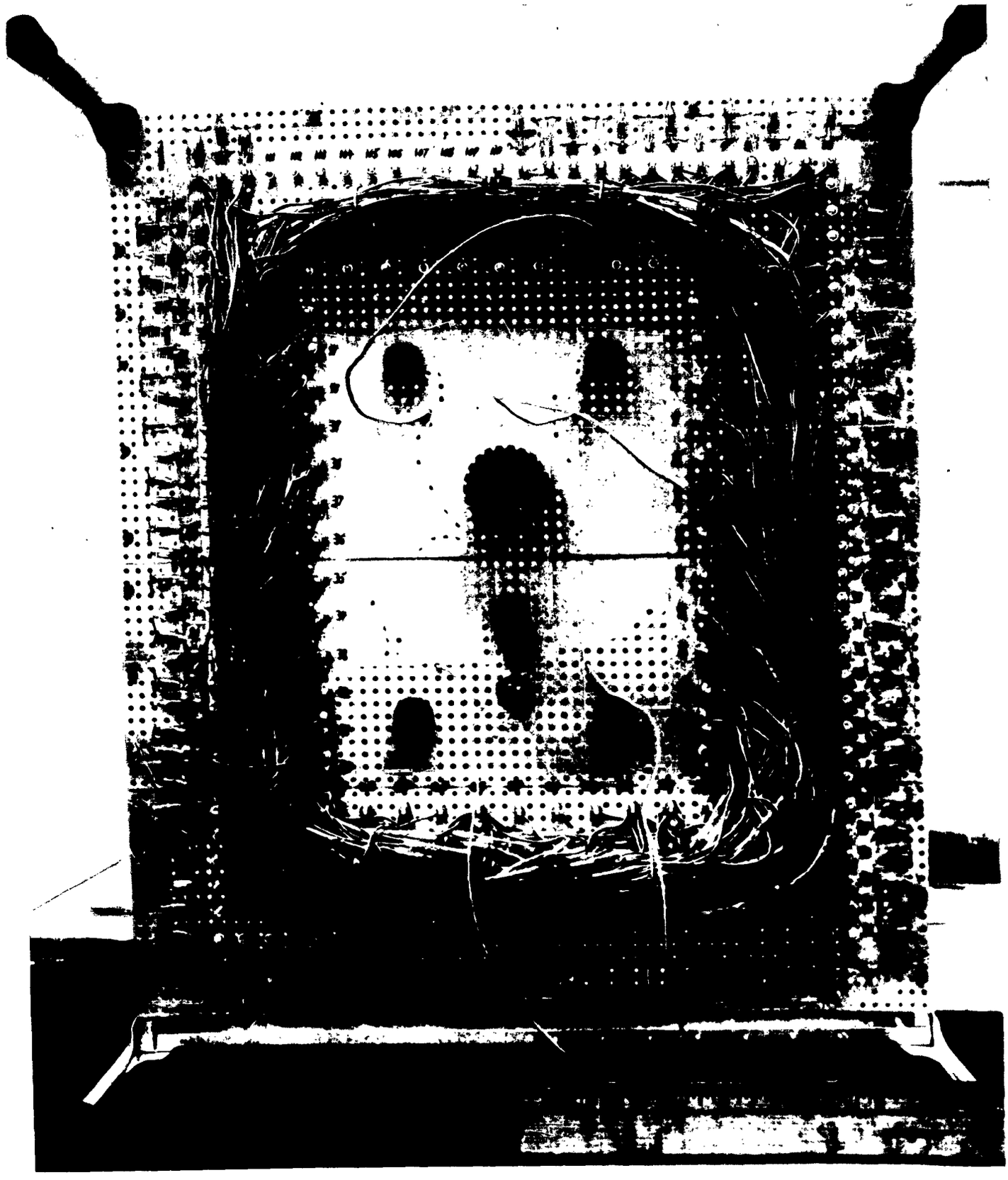
PARTIAL VOCABULARY OF KEY TERMS FOR THE

ACORN-I DEMONSTRATION

1. Air Force
2. Nuclear Warhead
3. Military Establishment
4. Missile Interception
5. Air Defense
6. Aircraft
7. Satellite Missiles
8. Command and Control
9. Navy
10. Intermediate Range Missiles
11. Army
12. ICBM, ICBM's
13. Surveillance
14. Tactical Missions
15. Strategic Deterrence
16. SAC
17. National Security
18. NATO
19. Over-Water
20. Soviets
21. Continental Defence
22. Transonic Bomber
23. Over-land
24. Submarines



ACORN-1: (Associative Content Retrieval Network) Wired for 40 sentences using 40 index terms. To pose an inquiry, the wires with the clips are attached to the terminals for the index terms and/or sentences deemed to be relevant by the user. As the large knob is turned the voltages on these wires are raised, and the neon bulbs light up in the order of "relevance" of the various sentences. Relative voltages on the individual wires are controlled by the other knobs. Association may be set either "free" or "narrow" by varying the setting on the lower right hand knob.



ACORN-2: (Associative Content Retrieval Network) Wired for 100 libraries covering 42 subject areas. To pose an inquiry, the wires with the clips are attached to the terminals for the topics and/or libraries deemed to be relevant by the user. As the large knob is turned the voltages on these wires are raised, and the neon bulbs light up in the order of "relevance" of the various libraries. Relative voltages on the individual wires are controlled by the other knobs. Association may be set either "free" or "narrow" by varying the setting on the lower center knob.